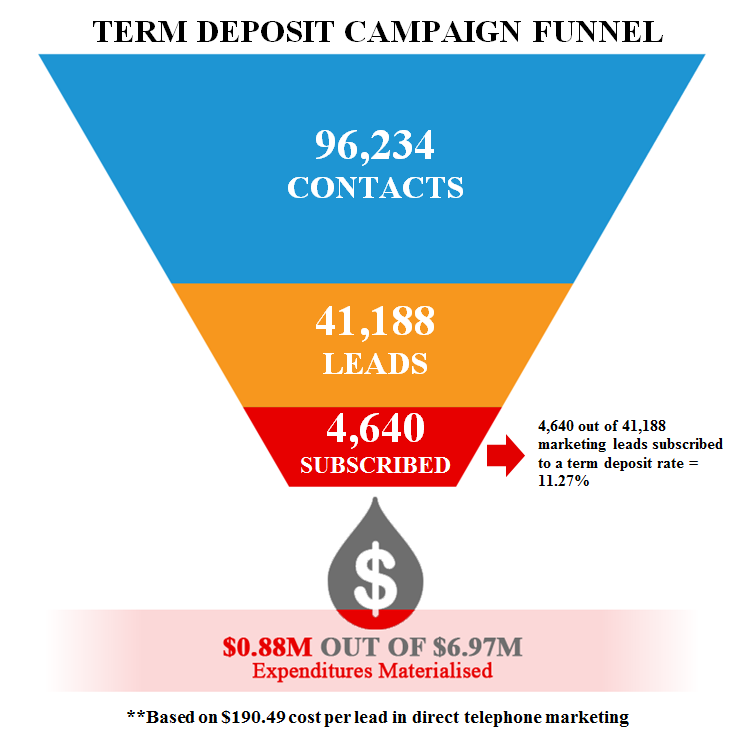
**COMP5310: Binary Classification Model to Predict Term Deposit Campaign Subscription**

Emil Laurence Pastor

epas6415@uni.sydney.edu.au

The University of Sydney

**Executive Summary**



According to the *Gartner 2016-2017 CMO Spend Survey,* marketing budgets continued their steady ascent in 2016, climbing to 12% of company revenue(Pemberton, 2017). Do millions of dollars in marketing expenditure can really bring the value to the company and their customers? A marketing data suggest that the average lead to sales conversion rate in upselling a product or service to an existing customer is just 11.27% in a typical direct marketing campaign. Furthermore, this means that around 88.73% of the expenditures didn’t translate to actual sales. Moreover, the measly lead to sales conversion rate and unnecessary operational expenditures is the result of (1) poor market segmentation, (2) channel selection, (3) customers receiving several campaigns over a period and (4) not taking advantage of the customer and transactional data. In line with this, companies should re-evaluate their current marketing strategy on how they can best use their existing assets to improve their KPIs.

Figure 1 Term Deposit Campaign Funnel

To overcome these challenges, companies like banks should monetise on the immense amount of customer and transactional data available in their systems to improve their marketing strategy to a data driven and highly targeted analytical campaign. Having said that, this project will take advantage years’ worth of data to build a binary classification model that will predict if the lead will subscribe to a product or service. Thus, this approach aims to translate into an increase in the lead to sales conversion rate, increase operational efficiencies, improved customer segmentation and better customer satisfaction through the reduction of unnecessary contact while learning more about your customer.

**I. Problem Statement**

C

ompanies like banks spend millions of dollars on marketing which includes advertisements, promotions, and campaigns. Consequently, companies should re-evaluate their expenditure and strategy alongside with the result marketing KPIs like lead to sales conversion rate. In line with this, the data gathered between May 2008 to November 2010 from a European bank show that telemarketers contacted the leads 2.57 times to have 4,640 out of 41,188 marketing leads subscribed to a term deposit direct marketing campaign. Furthermore, this result in a low lead to sales conversion rate of 11.27% and $0.88M out of $6.97M expenditure translated into sales using $190.49 cost per lead in direct telephone marketing (Lohrey, 2013). Thus, this scenario creates a long-term and adverse problem resulting to poor customer engagement and satisfaction, decrease in operational efficiencies and low marketing KPIs like lead to sales conversion rate. A successful solution is to build a binary classification model that label customers who will be part of the campaign leads to be called by telemarketers for a term deposit subscription offer. To be able to meet the business requirements, this study aims to answer the following challenges:

**Main Problem:**

***H0***: Multinomial Naïve Bayes and Logistic Regression shows no difference in the classification performance.

***H1***: Logistic Regression classifier performs better than Multinomial Naïve Bayes classifier.

**Supplementary Problems:**

1. Determine the features’ score in the data set.
2. Evaluate the appropriate method in handling an imbalanced data set.

**II. Data Set**

**1. General Information**

In this project, the Bank Marketing Data from *kaggle.com* and the *UCI Machine Learning repository* was used as the data sets for data exploration and analysis. The zip file consists of two *csv* files: (1) *bank-additional-full.csv* contains 41,188 records is used as training data set for the next stage of the project and (2) *bank-additional.csv* contains 4,119 records is used as the test data. Furthermore, the data set consists of 1 output variable (campaign outcome *y*) and 15 categorical and numeric features can be grouped into 3 categories: Customer Demographics, Customer Bank Information, and Historical Marketing Activity. The mix of different data categories provides a broader perspective in understanding the marketing bank and how does each feature contributes on the overall campaign outcome.

Table 1 Features and Output Data

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Data Type** |
| **Customer Demographics** | | |
| Age | Customer Age | Numeric |
| Job | Job type (Administration, Blue Collar, Management, etc.) | Categorical |
| Marital Status | Marital Status (Single, Divorced, Married) | Categorical |
| Educational Attainment | Highest Educational Attainment (High School, University Degree) | Categorical |
| **Customer Bank Information** | | |
| Credit Default Flag | Flag if the customer has credit in default | Categorical |
| Housing Loan Flag | Flag if the customer has housing loan | Categorical |
| Personal Loan Flag | Flag if the customer has personal loan | Categorical |
| **Historical Marketing Activity** | | |
| Contact Channel Type | Marketing contact channel type | Categorical |
| Contact Month | Last contact month | Categorical |
| Contact Day of Week | Last contact day | Categorical |
| Last Contact Duration | Last contact duration in seconds | Numeric |
| Campaign Contacts | Number of contacts performed during this campaign and for this customer | Numeric |
| Previous Campaign Last Contact | Number of days that passed by after the customer was last contacted from a previous campaign | Numeric |
| Previous Campaign Contact | Number of contacts performed before this campaign | Numeric |
| Previous Campaign Outcome | Outcome of the previous marketing campaign | Categorical |
| **Current Campaign Result** | | |
| Campaign Outcome | Term Deposit campaign outcome | Categorical |



Figure 2 Logical Data Model

In data management perspective, the data has been modelled into Star Schema to conform and organise the data set before performing further data deep dives and analysis as shown in *Figure 2 Logical Data Model*. The categorical variables are converted into a dimension which consists of source key, numerical target for labelling and target categorical value.

**2. Data Preparation and Exploration**



Figure 3 Data Transformations

Data preparation involves series of cleansing and transformation stages in the data as shown in *Figure 3 Data Transformations* to ensure that there are no missing values and the presentation layer is compatible to the classification algorithms. Having said that, one hot encoding technique is used for categorical variables where each unique category is represented as a column (*Figure 4 One Hot Encoding Process*). This feature representation works better with classification algorithms.



Figure 4 One Hot Encoding Process

The next step is to explore the data set to generate insights to further understand the data. *Figure 5* Campaign Outcome Distribution shows that the campaign outcome has an imbalanced distribution between the Yes and No. Thus, strategy to handle imbalanced data is required prior to building the model.

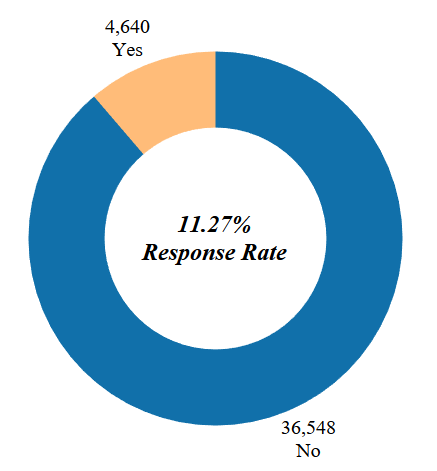


Figure 5 Campaign Outcome Distribution

**III. Quantitative Evaluation Methods**

To be able to meet the business requirements, this study aims to compare the performance of the two classification algorithms that can be used to build the binary classification model. Furthermore, the metrics to be used must take into consideration that the data is imbalanced with the goal to have a good balance between the precision and recall. Consequently, the following metrics are used for classifier performance evaluation:

**1. f1-score**

This is the harmonic mean of precision and recall which conveys the balance between the two parameters. In addition, this measure provides an indication of how well classifier in confirming that the number of misclassified customers is at minimum while maximising the number of relevant customers.

**2. Matthews Correlation Coefficient**

MCC is used to measure the quality of binary classifiers. It considers true and false positives and negatives and generally regarded as a balanced measure which can be used even if the classes are imbalanced.

* *TP:* True Positive
* *FN:* False Negative
* *FP:* False Positive
* *TN:* True Negative

**Statistical Significance Test**

The Wilcoxon signed-rank test is the non-parametric test used to determine the statistical significance in accepting or rejecting the null hypothesis H0 to assess whether the two measurement populations have the same distribution. For this study, the confidence interval is set to α = 95%.

**IV. Experimental Setup**

**1. Architecture**

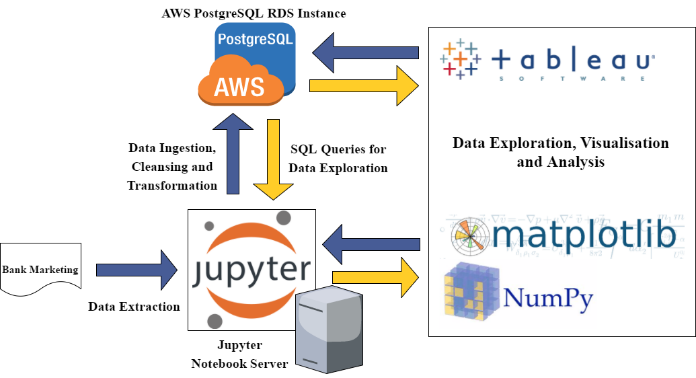


Figure 6 Architecture

The Data Architecture as shown in *Figure 6 Architecture* provides a high-level strategy and tools used for this stage of the project. Having said that, *Jupyter Notebook* is used for compiling and execution of *Python 3.6* codes with the use of *matplotlib* for visualisation, and *numpy* for statistical and numerical calculations, the *AWS PostgreSQL* instance was created to stage, cleanse and transform data to fit into the data model and Tableau was used for most of the data visualisation activities.

**2. Experimental Approach**



Figure 7 Experimental Approach

In this experiment, the approach is start with feature scoring using score function *f\_classif*. This step also selects the top 25 (after hot encoding process) feature based on the f-score. Subsequently, methods of handling imbalanced: 1.) Random Over-sampling, 2.) Random Under-sampling, 3.) Edited Nearest Neighbour, 4.) Synthetic Minority Over-Sampling Technique (SMOTE) is performed and verified against the default Multinomial Naïve Bayes and Logistic Regression algorithm. Then, the best performing imbalanced handling method is used for the subsequent step which is model tuning and parameter selection of Multinomial Naïve Bayes and Logistic Regression. This step uses the Grid Search function which tests the different supplied parameter and select the best possible parameters based on the f1-score of the individual models. The table below shows the parameters used for tuning the algorithm parameters:

Table 2 Algorithm Parameters for Tuning

|  |  |
| --- | --- |
| **Multinomial Naïve Bayes** | **Logistic Regression** |
| alpha: [0.01, 0.1, 1.0, 10.0, 100.0] | C: [0.01, 0.1, 1.0, 10.0, 100.0] |
|  | penalty: [ L1, L2] |
|  | class weight: [None, balanced] |
|  | max iteration: [10, 100, 1000] |

The best parameters for Multinomial Naïve Bayes is used as the benchmark and the performance comparison with Logistic Regression for model evaluation and performance analysis.



Figure 8 Bootstrap Method

Bootstrap method is used as a method of testing the hypothesis statistical significance. Moreover, the approach is to combine the training and testing set and conduct 50 experiments of randomly generated 80% training set and 20% testing set which subsequently generates measurement population of f1-score and MCC from Multinomial Naïve Bayes and Logistic Regression classification model. Furthermore, the generated result is used in performing Wilcoxon signed-rank test. Compared to basic cross-validation, the bootstrap method increases the variance that can occur in each fold. This is a desirable property since it is a more realistic simulation of the real-life experiment from which the dataset is obtained (Dougherty, 2013).

Further analysis is conducted in this experiment to provide more insights on the performance of the classifier and translate it back to the business requirements. Having said that, the following evaluation methodology is also used:

1. Confusion Matrix
2. Receiver Operating Characteristic (ROC)
3. Precision-Recall Curve

**V. Results**

**1. Feature Analysis**

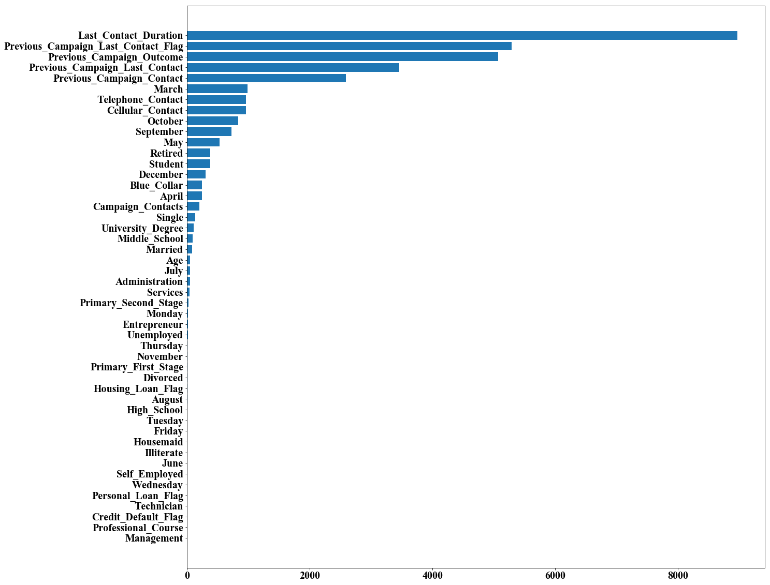


Figure 9 SelectKBest Feature Score in descending order

The feature analysis using SelectKBest f\_classif function (*Figure 9* SelectKBest Feature Score) shows that the top 5 best scoring features are historical marketing activity of the customer with the Last Contact Duration as the top scoring feature. This implies that historical information about the customer marketing activity tend to influence the overall result of future marketing activity.

Table 3 Effect of Top 5 Scoring Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measures** | **Multinomial Naïve Bayes** | | **Logistic Regression** | |
|  | **All** | **Without Top 5** | **All** | **Without Top 5** |
| f1-score (Overall) | 0.8700 | 0.8400 | 0.9000 | 0.8600 |
| f1-score  (Class Yes) | 0.4400 | 0.2600 | 0.4600 | 0.1500 |
| MCC | 0.3735 | 0.2456 | 0.4442 | 0.1844 |

Subtractive feature analysis is also conducted to verify the effect in the metrics if the Top 5 scoring features are removed. *Table 3* Effect of Top 5 Scoring Featuresshows that removing the top 5 scoring feature resulted to drop in the measures for both algorithms except for the overall f1-score in Multinomial Naïve Bayes.

**2. Handling Imbalanced Data Set**

The distribution of the labels is skewed towards the class No resulting to imbalanced data set. This scenario is resulting to poor measures of the minority class like f1-score and Matthews Correlation Coefficient (MCC) as shown in *Table 4* Methods to Handle Imbalanced Data. Different methods are applied to improve the performance of the imbalanced data set. Having said that, the Random Over-sampling method provides the most improvement in terms of the minority class f1-score and Matthews Correlation Coefficient (MCC) in Logistic Regression while not much difference in Multinomial Naïve Bayes. Therefore, the training data is transformed using Random Over-sampling and used for further processing.

Table 4 Methods to Handle Imbalanced Data

|  |  |  |
| --- | --- | --- |
| **Methods** | **Multinomial Naïve Bayes** | **Logistic Regression** |
| **Overall f1-score** | | |
| No Method | 0.8400 | 0.9000 |
| Random Over-sampling | 0.8400 | 0.9000 |
| Random Under-sampling | 0.8300 | 0.8900 |
| Edited Nearest Neighbour | 0.8000 | 0.8900 |
| SMOTE | 0.8300 | 0.8900 |
| **Minority Class (Yes) f1-score** | | |
| No Method | 0.4400 | 0.4600 |
| Random Over-sampling | 0.4300 | 0.5700 |
| Random Under-sampling | 0.4400 | 0.5800 |
| Edited Nearest Neighbour | 0.4100 | 0.5600 |
| SMOTE | 0.4300 | 0.5700 |
| **Matthews Correlation Coefficient** | | |
| No Method | 0.3704 | 0.4426 |
| Random Over-sampling | 0.3724 | 0.5224 |
| Random Under-sampling | 0.3698 | 0.5189 |
| Edited Nearest Neighbour | 0.3572 | 0.5083 |
| SMOTE | 0.3671 | 0.5223 |

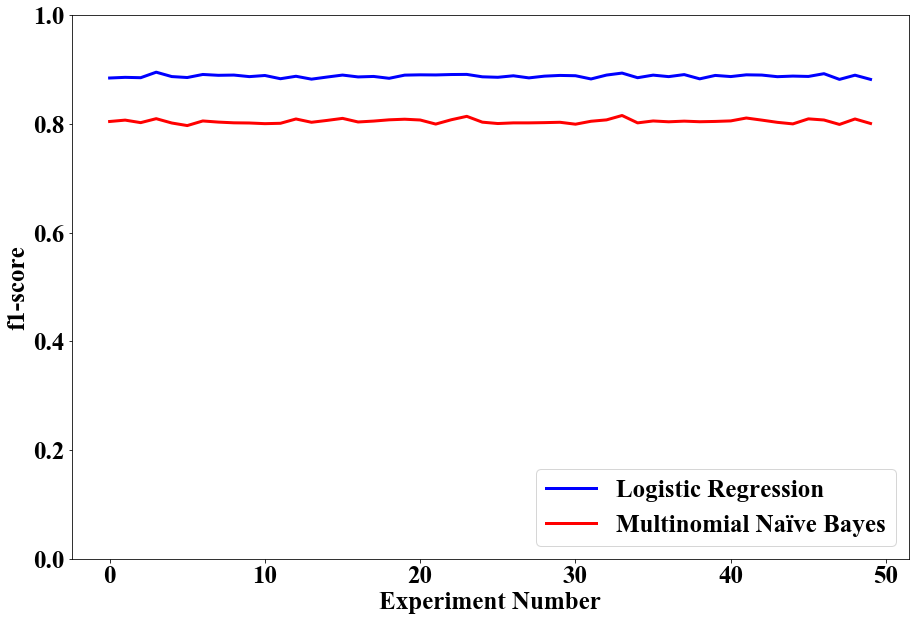
**3. Model Tuning and Parameter Selection**

Table 5 Best Model Parameters

|  |  |
| --- | --- |
| **Multinomial Naïve Bayes** | |
| **Parameter** | **Value** |
| alpha smoothing parameter | 0.01 |
| **Logistic Regression** | |
| **Parameter** | **Value** |
| C inverse regularisation strength | 1 |
| class weight | None |
| maximum iteration | 100 |
| Penalty | L2 |

Grid Search CV is used to optimise the Multinomial Naïve Bayes and Logistic Regression model ensuring optimal values based on the f1-score. The Table 5 Best Model Parameters below are the best parameters to be used on the respective classification algorithms.

**4. Summary**



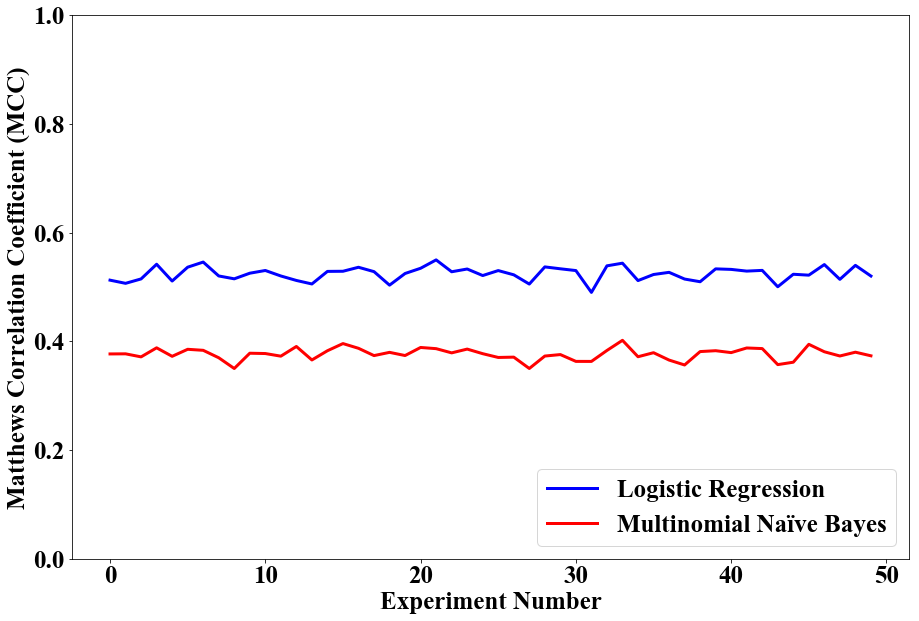


Figure 10 f1-score and MCC Measurement Population

The f1-score and MCC population measure is generated after performing the Bootstrap method using 50 experiments as shown in *Figure 10* f1-score and MCC Measurement Population. In addition, the training data in the Bootstrap is also transformed using Random Over-sampling method. After performing the experiments, the p-value for f1-score is 7.5518 x 10-10 and MCC is 7.5569 x 10-10 at 95% confidence interval using Wilcoxon signed-rank test. Thus, the null hypothesis that both classifier does not have significant performance difference is rejected. Refer to *Table 6* Wilcoxon Signed-rank Test Result.

Table 6 Wilcoxon Signed-rank Test Result

|  |  |  |
| --- | --- | --- |
| **Measure** | **p-value** | **Result** |
| f1-score | 7.5518 x 10-10 | Reject H0 |
| Matthews Correlation Coefficient (MCC) | 7.5569 x 10-10 | Reject H0 |

The ***Error! Not a valid bookmark self-reference.*** shows the summary of different measures of the two classifiers. The recall of the minority class is the same for both model while there is observable difference on their precision. Moreover, the Logistic Regression model has better performance in all the measures except for slight difference in recall.

Table 7 Summary of Results

|  |  |  |
| --- | --- | --- |
| **Measure** | **Multinomial Naïve Bayes** | **Logistic Regression** |
| **Accuracy** | | |
| Overall | 0.8307 | 0.8893 |
| **Precision** | | |
| Class Yes | 0.3200 | 0.5000 |
| Class No | 0.9600 | 0.9600 |
| Overall | 0.8900 | 0.9100 |
| **Recall** | | |
| Class Yes | 0.6900 | 0.6900 |
| Class No | 0.8200 | 0.9100 |
| Overall | 0.8000 | 0.8900 |
| **f1-score** | | |
| Class Yes | 0.4400 | 0.5800 |
| Class No | 0.8800 | 0.9400 |
| Overall | 0.8300 | 0.9000 |
| **Matthews Correlation Coefficient** | | |
| Overall | 0.3724 | 0.5255 |

Table 8 Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **Multinomial Naïve Bayes** | | | |
|  | **Predicted** | | |
| **Actual** | **Class** | **Yes** | **No** |
| **Yes** | 310 | 141 |
| **No** | 663 | 3005 |
| **Logistic Regression** | | | |
|  | **Predicted** | | |
| **Actual** | **Class** | **Yes** | **No** |
| **Yes** | 312 | 139 |
| **No** | 317 | 3351 |

The confusion matrix of both classifier shows both classifier have comparable lost opportunities (False Negatives) in the minority class. Nevertheless, Logistic Regression has better precision at 50% compared to Multinomial Naïve Bayes with 32%. In the business sense, there are more potential Return on Investment (ROI) in LR than MNB. Overall, both classifiers perform better than the original response rate of 11.27% should the predicted class Yes be used as the new pool of leads for the campaign.

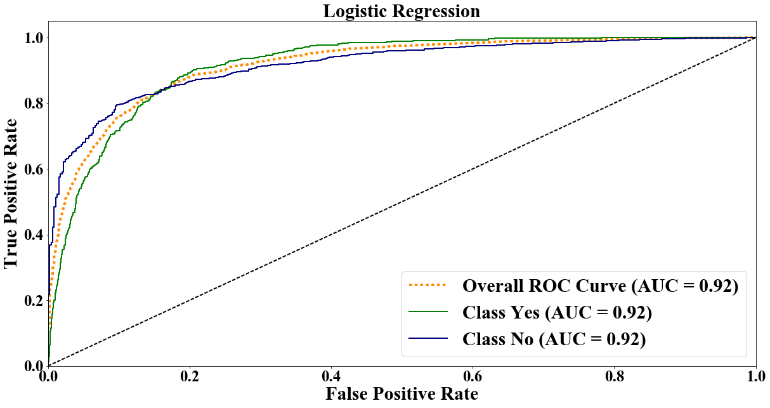
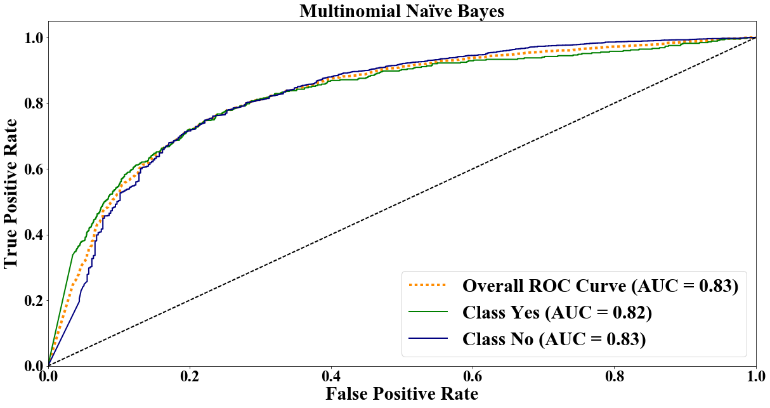
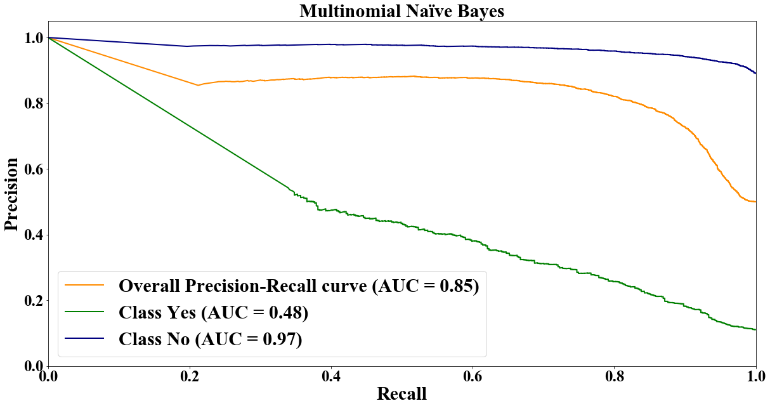


Figure 11 Receiver Operating Characteristic (ROC)

On the other hand, the Receiver Operating Characteristic (ROC) curve is visually appealing to provide an overview the classifier’s performance across a wide range of specificities (Rehmsmeier, M. and Saito, T, 2015). While it shows that the ROC Area Under the Curve performance of LR is better than MNB, it does not tell enough insights about the performance of the minority class as shown in *Figure 11* Receiver Operating Characteristic (ROC).

As an alternative, the figures below provide clearer visual indications that allows an accurate and intuitive interpretation of the classifier performance. Having said that, the majority class has a better PRC over the minority class for both classifier. However, the PRC of Logistic Regression is better than the Multinomial Naïve Bayes with Area Under the Curve of 0.57 compare to 0.48.

 Figure 12 Precision-Recall Curve (PRC) for Multinomial Naïve Bayes

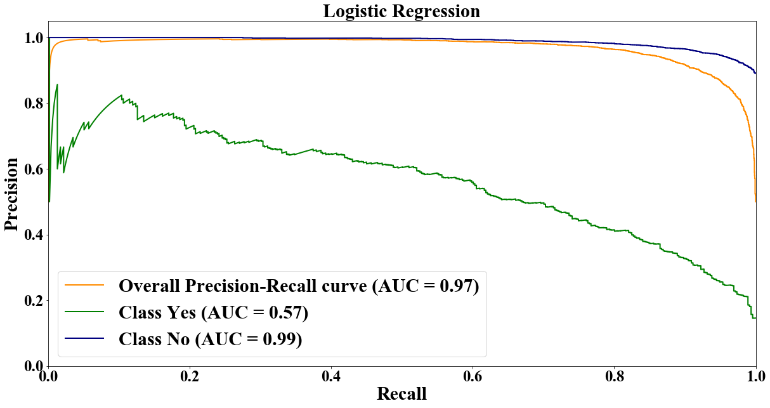


Figure 13 Precision-Recall Curve (PRC) for Logistic Regression

**5. Future Work and Learnings**

This study meets the business objective of identifying the potential leads who will subscribed to the term deposit and the results shows considerable improvement in the response should the Logistic Regression be used. However, there are room for improvement in this study which the future proponents can undertake:

* Leverage more historical banking and marketing activity features like the time of the day, dates of previous contact, last bank related transaction and the type of transactions.
* Use different classification algorithm like Random Forest Classifier and verify any improvements in the performance.
* Anomaly and outlier detection can also be a potential future that can further improve the overall measures.
* Look at the Odds-ratio of response perspective.

This study provides significant learnings to the author in dealing with a real-life data science project primarily on the role of data scientist to bridge the gap between data and relevant business objective using software development, statistics, and business intelligence to maximise the business output. Furthermore, the author learned the importance of each phase of the development particularly business and data understanding, data modelling and feature selection, model selection and parameter tuning to have a successful result.

**VI. Conclusion**

Overall, the solution is successful in satisfying the business objective of identifying potential pool of leads from the test data and accepting the alternative hypothesis that Logistic Regression performs better based on the f1-score and Matthew’s correlation coefficient than Multinomial Naïve Bayes at 95% confidence interval. Moreover, the response rate of the chosen model: Logistic Regression (using the customer that is classified as Class Yes to be the leads) delivers a 50% response rate compared to the original response rate of 11.27%. This result translates into an increase in the lead to sales conversion rate, increase in operational efficiencies, improvement in customer segmentation and better customer satisfaction since unnecessary contact is minimised.

**VI. References**

Brownlee, J (2014, October 4) An Introduction to Feature Selection. Retrieved from http://machinelearningmastery.com/an-introduction-to-feature-selection/.

CRISP-DM Consortium (2000, August). CRISP-DM 1.0 Step-by-step Data Mining Guide. Retrieved from http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf.

Dougherty, G. (2013). Pattern Recognition and Classification: An Introduction. pp 157-171.

Lohrey, J. (2013). The Average Success Rate of Direct Marketing. Retrieved from http://smallbusiness.chron.com/average-success-rate-direct-marketing-73648.html

Markham, Kevin (2014, March 25) Simple guide to confusion matrix terminology. Retrieved from http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/.

Moorman, C. and Finch, T.A. (2017, January 24). Marketing Budgets Vary by Industry. Retrieved from http://deloitte.wsj.com/cmo/2017/01/24/who-has-the-biggest-marketing-budgets/

Moro, S. (2014) Bank Marketing Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Pemberton, C. (2016, December 12). Gartner CMO Spend Survey 2016-2017 Shows Marketing Budgets Continue to Climb. Retrieved from http://www.gartner.com/smarterwithgartner/gartner-cmo-spend-survey-2016-2017-shows-marketing-budgets-continue-to-climb/

Rehmsmeier, M. and Saito, T (2015, March 4). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.